Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati

Department of Artificial Intelligence and Data Science

S.Y. B. Tech Syllabus 2024-25 (As per NEP 2020)

Syllabus: Multidisciplinary Minor
w. e. f. AY: 2024- 2025
SEMESTER-IV

Multidisciplinary Minor in Artificial Intelligence and Data Science

SEM	Course Code	Courses Name	Teaching Scheme			Examination Scheme and Marks								Credits		
			TH	PR	TUT	Activity	ISE	ESE	TW	PR	OR	Total	TH	PR	TUT	Total
IV	AI23051	Data Processing and Analysis	2	2	-	20	20	50	20			110	2	1		3

Pshendage

Dept. Academic Coordinator

Mr. P.N. Shendage

Head of Department

Dr. C. S. Kulkarni

02/01/2025

Dean Academic

Dr. S. M. Bhosle

Principal

Dr. S. B. Lande

BUCKET OF MULTIDISCIPLINARY SUBJECT

MULTIDISCIPLINARY MINOR SUBJECT (only for students having CGPA >= 7.5)
AI23051: Data Processing and Analysis

VPKBIET BARAMATI
Affiliated to S.P.P.U. Pune Id No. PU/PN/Engg./152/2000
PUNE (MS.)

Vidya Pratishthan's

Kamalnayan Bajaj Institute of Engineering and Technology, Baramati

(Autonomous Institute)

AI23051- Data Processing and Analysis

Teaching Scheme: **Theory: 2 Hours/Week** **Practical: 2 Hour/Week**	**Credits** **03**	**Examination Scheme:** **Activity:20 Marks** **ISE: 20 Marks** **ESE: 50 Marks** **Term Work: 20 Marks**

Prerequisites: Python Programming

Course Objectives:

- To understand the need of Data Science
- To understand computational statistics in Data Science
- To provide a comprehensive knowledge of data science using Python.
- To learn the essential concepts of data analytics and data visualization.

Course Outcomes (COs): The students will be able to learn:

CO1: Apply basic data manipulation techniques using Pandas.
CO2: Identify and apply the need and importance of pre-processing techniques.
CO3: Implement data visualization using visualization tools in Python programming.
CO4: Apply various algorithms and evaluate their performance using metrics.

Course Contents

Unit I Introduction to Data Science and Pandas Basics (06 Hours)

Data science: Definition and importance of data science in various industries. Overview of the data science process and the role of a data scientist. Datafication, and data science lifecycle. Ethical considerations and challenges in data science.

Getting Started with Pandas: Overview of Pandas library and its architecture. Introduction to data structures: Series and DataFrames. Indexing, selection, and filtering data. Basic operations: sorting, ranking, reindexing, and handling missing data.

Unit II Statistical Inference & Data Wrangling (6 Hours)

Statistical Inference: Measures of central tendency: Mean, median, mode, and their application in data analysis. Measures of dispersion: Variance, standard deviation, and range. Introduction to Bayes' Theorem and its relevance to data science. Pearson Correlation and its role in understanding relationships between variables.

Data Wrangling: Combining and merging datasets using Pandas. Techniques for reshaping data (pivoting, melting). Handling overlapping data, removing duplicates, and replacing values. Transforming data for analysis (scaling, encoding, and feature extraction).

Unit III Plotting, Visualization & Exploratory Data Analysis (EDA) (6 Hours)

Plotting & Visualization: Importance of data visualization in data science. Overview of different types of data visualizations: Line plots, bar charts, histograms, scatter plots. Introduction to **Matplotlib** and **Seaborn** for data visualization in Python. Plot customization: Titles, labels, colors, legends, and annotations.

Exploratory Data Analysis (EDA): Visualizing distributions, relationships, and trends in the data. Using heatmaps, pairplots, and correlation matrices for EDA. Identifying patterns, outliers, and potential data issues through visualization.

Unit IV Machine Learning Basics & Model Evaluation (6 Hours)

Introduction to Machine Learning: Overview of machine learning: Types of learning (supervised vs unsupervised). Introduction to classification and regression tasks. Clustering algorithms: K-Means and hierarchical clustering. Evaluation metrics: Accuracy, precision, recall, F1-score.

Model Evaluation & Selection: Cross-validation and its importance in model evaluation. Overfitting and underfitting: Understanding and mitigating these issues. Introduction to Ridge regression for regularization.

Text Books:

1. David Dietrich, Barry Hiller, "Data Science and Big Data Analytics", EMC Education services, Wiley publication, 2012, ISBN0-07-120413-X.
2. Wes McKinney, "Python for Data Analysis",O'REILLY, ISBN:978-1-449-31979-3. 1st edition, October 2012.
3. Rachel Schutt & O'neil, "Doing Data Science", O'REILLY, ISBN:978-1-449-35865-5, 1st edition, October 2013.

Reference Books:

1. Joel Grus, "Data Science from Scratch: First Principles with Python", O'Reilly Media, 2015
2. Matt Harrison, "Learning the Pandas Library: Python Tools for Data Munging, Analysis, and Visualization , O'Reilly, 2016.
3. Chirag Shah, "A Hands-On Introduction to Data Science", Cambridge University Press, (2020), ISBN: 978-1-108-47244-9.
4. Wes McKinney, "Python for Data Analysis", O'Reilly media, ISBN: 978-1-449-31979-3.

E-Resources:

- https://onlinecourses.nptel.ac.in/noc21_cs69/preview
- https://nptel.ac.in/courses/106106179
- https://onlinecourses.swayam2.ac.in/imb23_mg64/preview

List of Assignments

1. Perform the following operations using Python on any open source dataset (e.g., data.csv)
   - Import all the required Python Libraries.
   - Locate open source data from the web (e.g., https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site).
   - Load the Dataset into pandas dataframe.
   - Data Preprocessing: check for missing values in the data using pandas isnull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.
2. Load the Iris dataset as a list of lists (each of the 150 lists should have 5 elements). Compute and print the mean and the standard deviation for each of the 4 measurement columns (i.e. sepal length and width, petal length and width). Compute and print the mean and the standard deviation for each of the 4 measurement columns, separately for each of the three Iris species (Versicolor, Virginica and Setose). Which measurement would you consider "best", if you were to guess the Iris species based only on those four values?
3. Create various plots (line, bar, scatter, histogram) using **Matplotlib** and **Seaborn**. Perform EDA on a dataset and visualize distributions, relationships, and correlations.
4. Implement a simple machine learning model (e.g., K-Means clustering or ridge regression) and evaluate its performance using cross-validation.
5. Demonstrate the K-Means Clustering model and evaluate the performance on Iris dataset.

Vidya Pratishthan's

Kamalnayan Bajaj Institute of Engineering and Technology, Baramati

Department of Artificial Intelligence and Data Science

Plan for Activity (20 Marks)

Course Name: - Data Processing and Analysis

Course Code: AI23051

Year: - SY **Branch: - AI&DS**

Semester: - IV

The course **Data Processing and Analysis** at the second-year level, Semester IV of the Artificial Intelligence and Data Science program, includes the following evaluation scheme:

SEM	Course Code	Courses Name	Teaching Scheme			Examination Scheme and Marks							Credits			
			TH	PR	TUT	Activity	ISE	ESE	TW	PR	OR	Total	TH	PR	TUT	Total
IV	AI23051	Data Processing and Analysis	2	2	-	20	20	50	20			110	2	1		3

The evaluation under the "**Activity**" component, worth 20 marks, will consist of a **quiz** featuring Multiple Choice Questions from various categories. The distribution of questions will be as follows: 20% difficult questions (from the Evaluate and Create categories of Bloom's Taxonomy), 40% medium questions (from the Apply and Analyze categories), and 40% easy questions (from the Remember and Understand categories). The schedule for this activity will be communicated via email, noticeboard, or the department website well in advance.

Students who fail to attend this activity but have a genuine reason will be accommodated with a rescheduled quiz, which will also be announced through email, noticeboard, or the department website well in advance, using a different set of questions from the specified categories.

Gawade

Ms. Roshani R. Gawade

Course Coordinator

Dr. C. S. Kulkarni

HoD

VPKBIET BARAMATI
Affiliated to
S.P.P.U. Pune
Id No./PU/PN/
Engg./152/2000
PUNE (MS.)

Head
Department of Artificial Intelligence
& Data Science,
VPKBIET, Baramati 413 133